

# META-SONG EVALUATION FOR CHORD RECOGNITION

Yizhao Ni<sup>1</sup>, Matt Mcvicar<sup>1</sup>, Raul Santos-Rodriguez<sup>2</sup> and Tijl De Bie<sup>1</sup>

1. Intelligent Systems Lab  
Department of Engineering Mathematics  
University of Bristol, U. K.

2. Signal Theory and Communications Department  
Universidad Carlos III de Madrid  
Spain

## ABSTRACT

We present a new approach to evaluate chord recognition systems on songs which do not have full annotations. The principle is to use online chord databases to generate high accurate “pseudo annotations” for these songs and compute “pseudo accuracies” of test systems. Statistical models that model the relationship between “pseudo accuracy” and real performance are then applied to estimate test systems’ performance. The approach goes beyond the existing evaluation metrics, allowing us to carry out extensive analysis on chord recognition systems, such as their generalizations to different genres. In the experiments we applied this method to evaluate three state-of-the-art chord recognition systems, of which the results verified its reliability.

## 1. INTRODUCTION

In recent years, audio chord recognition has become a very active field [1–4, 9, 10] due to the increasing popularity of Music Information Retrieval (MIR) with applications using mid-level tonal features has established chord recognition as a useful and challenging task.

Generally speaking, chord recognition is a task of automatically detecting chord labels and boundaries from the audio of a musical piece. The process involves segmenting a song into a high time resolution sequence of windows (known as *frames*), after which machine learning techniques (e.g. Hidden Markov Models) are utilized to detect chord label for each frame, based on the features extracted and the local context. The chord predictions can then be evaluated via frame-wise accuracies, if the ground truth annotation of the song is available.

The annual MIREX (Music Information Retrieval Evaluation eXchange) competition<sup>1</sup> has a task dedicated to chord recognition, where participants attempt to generate chord predictions for a collection of songs. In the most recent competitions, the dataset used is a collection of Beatles, Queen and Zweieck songs, of which the *ground truth annotations* are available. Due to the limited amount of data, existing chord recognition systems (referred as *test*

*systems* in the paper) are usually trained and tested on the same songs, inevitably causing over-fitting on this dataset. Meanwhile, the evaluation is also heavily constrained by the simplicity of the data. For example, most of the songs in the dataset are from Rock genre, implying that the performance lacks generalization to other genres.

To resolve these problems, the simplest, but most costly and least scalable solution would be to obtain more fully annotated data, paying trained musicians to annotate new songs. Alternatively, we propose using a methodologically more challenging but cheaper and scalable approach: *meta-song evaluation*, which makes use of large and freely available online chord databases, such as *E-chords*<sup>2</sup> to help evaluate test systems. The principle is to automatically generate chord annotations for new songs of which the chord sequences are available on these databases. The songs and the generated annotations are then used to estimate test systems’ performance via statistical theories.

However, chord sequences from such databases are generally less directly usable than those produced by musicians, since exact timings of chords are absent and sometimes the chord sequences are affected by various types of errors and omissions. Hence, a system, referred to as *reference system* in the paper, is required to generate high accurate annotations from these untimed chord sequences. As demonstrated in our previous work [6], we have designed a variety of reference systems from which the generated annotations are more accurate than most of the existing chord recognition systems. We regard these high accurate, but not perfect annotations as *pseudo annotations*. In the rest of this paper, we will show how to make use of these pseudo annotations to comprehensively evaluate performances of different test systems.

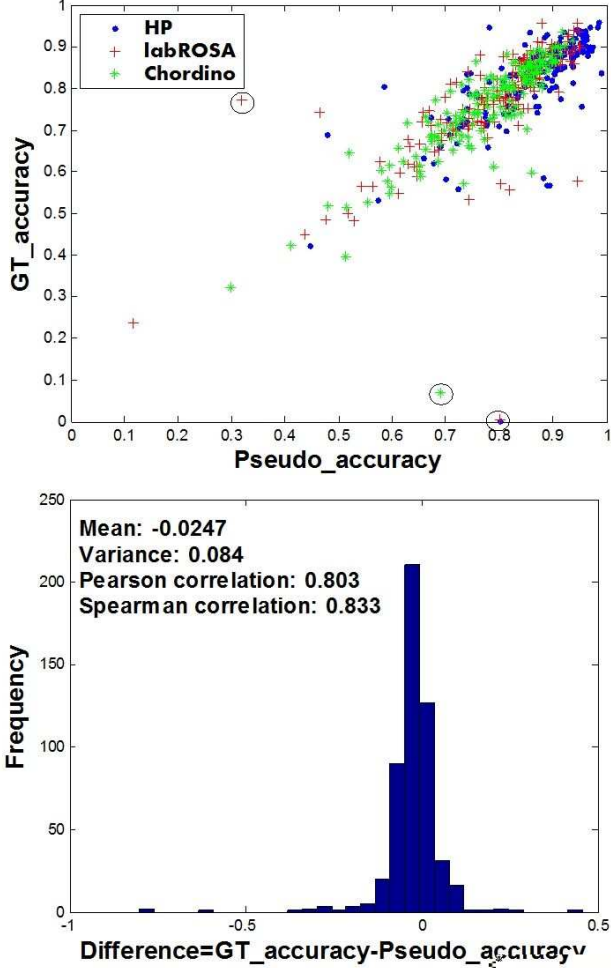
## 2. MATHEMATICAL FRAMEWORK

We use  $y_i^A$  and  $x_i^A$  to denote the *ground truth (GT) accuracy* and the *pseudo accuracy* (i.e. the accuracy of system’s prediction compared to pseudo annotation) of system  $A$ ’s chord prediction for the  $i$ -th song. Then for each system we obtain two sets of data: a validation set  $\{x_i^A, y_i^A\}_{i=1}^n$  and a test set  $\{x_j^A\}_{j=n+1}^{n+m}$ . Note that we only have ground truth annotations on the validation set such that generally  $m \gg n$ . The test system pool is denoted by  $A \in \mathcal{A}$ .

One observation from the validation set is that the pseudo accuracies are highly correlated with the GT accuracies

<sup>1</sup> [http://www.music-ir.org/mirex/wiki/MIREX\\_HOME](http://www.music-ir.org/mirex/wiki/MIREX_HOME)

<sup>2</sup> <http://www.e-chords.com/>



**Figure 1.** The relationship between the pseudo accuracies and the real GT accuracies on 175 The Beatles songs. The three systems are: A. HP, B. labROSA and C. Chordino. The pseudo annotations of the songs are generated by the Jump Alignment method [6], which has been shown to produce more accurate chord predictions than all other systems. Note that there are some outliers (represented by circled points on the top figure), of which the online chord sequences are less informative (e.g. the chord sequence only records the solo of the song). In these cases, the resulting pseudo accuracies are not well-correlated with the GT accuracies. How to reduce and eliminate these outliers will be investigated in our future work.

(see Figure 1), as long as the pseudo annotations are accurate enough. In the ideal case, if all pseudo annotations are 100% accurate, the pseudo accuracies will converge to the GT accuracies. Inspired by this observation, we propose three mathematical frameworks to model the relationship between GT and pseudo accuracies on the validation set, which can then be applied to estimate GT accuracies on the test set.

## 2.1 Single Gaussian model

This model assumes that the pairs  $(x_i, y_i)$  generated by all systems  $\mathcal{A}$  are sampled *i.i.d* from a single Gaussian distribution  $\mathcal{N}(\mu, \sigma^2)$ . That is

$$y_i^A = x_i^A + \mu + \epsilon_i, \quad 1 \leq i \leq n, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \forall A \in \mathcal{A}. \quad (1)$$

The parameters of the distribution  $\mu$  can then be easily estimated by *least square* [7], resulting in

$$\bar{\mu} = \frac{1}{|\mathcal{A}|n} \sum_{A \in \mathcal{A}} \sum_{i=1}^n (y_i^A - x_i^A).$$

The unbiased estimator of  $\sigma^2$  can be calculated by the formula

$$\bar{\sigma}^2 = \frac{1}{|\mathcal{A}|n - 1} \sum_{A \in \mathcal{A}} \sum_{i=1}^n (y_i^A - x_i^A - \bar{\mu})^2.$$

Using the parameters  $(\bar{\mu}, \bar{\sigma}^2)$  estimated from the validation set, we are able to predict the GT accuracies  $\{y_j^A\}_{j=n+1}^{n+m}$  on the test set using the linear regression theory. Let  $\bar{x} = \frac{1}{|\mathcal{A}|n} \sum_{A \in \mathcal{A}} \sum_{i=1}^n x_i^A$  and  $s_x^2 = \frac{1}{|\mathcal{A}|n - 1} \sum_{A \in \mathcal{A}} \sum_{i=1}^n (x_i^A - \bar{x})^2$ , the following Gaussian distribution holds for all test examples<sup>3</sup> [7]:

$$y_j^A - x_j^A - \bar{\mu} \sim \mathcal{N}\left(0, \bar{\sigma}^2 \left(1 + \frac{1}{|\mathcal{A}|n} + \frac{(x_j^A - \bar{x})^2}{(|\mathcal{A}|n - 1)s_x^2}\right)\right). \quad (2)$$

Therefore, with probability  $1 - \alpha$  the confidence interval of  $y_j^A$  is

$$y_j^A = x_j^A + \bar{\mu} \pm Q(1 - \alpha) \bar{\sigma} \left(1 + \frac{1}{|\mathcal{A}|n} + \frac{(x_j^A - \bar{x})^2}{(|\mathcal{A}|n - 1)s_x^2}\right)^{1/2}, \quad (3)$$

where  $Q$  denotes a normal quantile function  $Q(p) = \inf\{y \in \mathbb{R} : p \leq \Pr(Y \leq y)\}$ .

We then extend Eq. (2) and estimate the mean accuracy of the test set  $\bar{y}^A = \frac{1}{m} \sum_{j=n+1}^{n+m} y_j^A$  using

$$\bar{y}^A - \bar{x}^A - \bar{\mu} \sim \mathcal{N}(0, \hat{\sigma}_A^2),$$

$$\text{with } \bar{x}^A = \frac{1}{m} \sum_{j=n+1}^{n+m} x_j^A \text{ and } \hat{\sigma}_A^2 = \bar{\sigma}^2 \frac{\sum_{j=n+1}^{n+m} 1 + \frac{1}{|\mathcal{A}|n} + \frac{(x_j^A - \bar{x})^2}{(|\mathcal{A}|n - 1)s_x^2}}{m^2}.$$

Again with probability  $1 - \alpha$  the confidence interval of  $\bar{y}^A$

<sup>3</sup> From a purely probabilistic perspective, the test examples follow a student-t distribution instead of a Gaussian. But since  $n$  is large enough ( $n > 100$ ) in our case, we approximate the student-t distribution as a Gaussian.

is

$$\bar{y}^A = \bar{x}^A + \bar{\mu} \pm Q(1 - \alpha)\hat{\sigma}_A. \quad (4)$$

Apart from estimating the confidence interval of the GT accuracies, the Gaussian distribution also allows us to compare two systems  $A$  and  $B$ , by means of estimating the confidence interval of  $\bar{y}^A - \bar{y}^B$  using

$$\bar{y}^A - \bar{y}^B \sim \mathcal{N}(\bar{x}^A - \bar{x}^B, \hat{\sigma}_A^2 + \hat{\sigma}_B^2).$$

This yields the following confidence interval with probability  $1 - \alpha$

$$\bar{y}^A - \bar{y}^B = \bar{x}^A - \bar{x}^B \pm Q(1 - \alpha)\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}. \quad (5)$$

The advantage of the single Gaussian model is that it makes use of  $|\mathcal{A}|$  times data to estimate the Gaussian parameters, which is expected to provide more robust estimation. However, as we observed in the experiments, test systems that are closer to the reference system generally got higher pseudo accuracies than the others. In this case, the GT accuracies estimated by the single Gaussian model would bias towards these systems.

## 2.2 Individual Gaussian model

To reduce or eliminate the effect of such biases, we proposed a variant of the single Gaussian model, fitting individual Gaussians to different test systems. Mathematically, the GT accuracy  $y_i^A$  is now modelled as

$$y_i^A = x_i^A + \mu_A + \epsilon_i, \quad 1 \leq i \leq n, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_A^2), \quad \forall A \in \mathcal{A}, \quad (6)$$

where the parameters  $(\mu_A, \sigma_A^2)$  can be learnt from the validation data  $\{x_i^A, y_i^A\}_{i=1}^n$ :

$$\begin{cases} \bar{\mu}_A = \frac{1}{n} \sum_{i=1}^n (y_i^A - x_i^A) \\ \bar{\sigma}_A^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i^A - x_i^A - \bar{\mu})^2 \end{cases}, \quad \forall A \in \mathcal{A}. \quad (7)$$

Here we denote  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i^A$  and  $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i^A - \bar{x})^2$ , then following the same procedure as described in Section 2.1 we obtain

$$y_j^A - x_j^A - \bar{\mu}_A \sim \mathcal{N}\left(0, \bar{\sigma}_A^2 \left(1 + \frac{1}{n} + \frac{(x_j^A - \bar{x})^2}{(n-1)s_x^2}\right)\right),$$

and with probability  $1 - \alpha$  the confidence interval of  $y_j^A$  is

$$y_j^A = x_j^A + \bar{\mu}_A \pm Q(1 - \alpha)\bar{\sigma}_A \left(1 + \frac{1}{n} + \frac{(x_j^A - \bar{x})^2}{(n-1)s_x^2}\right)^{1/2}.$$

Similarly, we have

$$\bar{y}^A = \bar{x}^A + \bar{\mu} \pm Q(1 - \alpha)\hat{\sigma}_A, \quad (8)$$

$$\text{with } \bar{x}^A = \frac{1}{m} \sum_{j=n+1}^{n+m} x_j^A \text{ and } \hat{\sigma}_A^2 = \bar{\sigma}_A^2 \frac{\sum_{j=n+1}^{n+m} 1 + \frac{1}{n} + \frac{(x_j^A - \bar{x})^2}{(n-1)s_x^2}}{m^2}.$$

To compare the two systems  $A$  and  $B$ , we now have a term  $\bar{\mu}_A - \bar{\mu}_B$  to reduce the effect of the biases, yielding

$$\bar{y}^A - \bar{y}^B \sim \mathcal{N}(\bar{x}^A - \bar{x}^B + \bar{\mu}_A - \bar{\mu}_B, \hat{\sigma}_A^2 + \hat{\sigma}_B^2).$$

We derive the following confidence interval with probability  $1 - \alpha$

$$\bar{y}^A - \bar{y}^B = \bar{x}^A - \bar{x}^B + \bar{\mu}_A - \bar{\mu}_B \pm Q(1 - \alpha)\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}. \quad (9)$$

## 2.3 Linear regression model

Apart from applying different  $\mu_A$  to eliminate the biases, one can also learn the slope of the regression line to better fit the validation samples. Mathematically, the relationship between  $y_i^A$  and  $x_i^A$  is now formulated as

$$y_i^A = a_A x_i^A + b_A + \epsilon_i, \quad 1 \leq i \leq n, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_A^2), \quad \forall A \in \mathcal{A}, \quad (10)$$

where the parameters  $(a_A, b_A, \sigma_A^2)$  can be estimated by least square:  $\forall A \in \mathcal{A}$

$$\begin{cases} \bar{a}_A = \frac{\sum_{i=1}^n (y_i^A - \bar{y})(x_i^A - \bar{x})}{\sum_{i=1}^n (x_i^A - \bar{x})^2} \\ \bar{b}_A = \bar{y} - \bar{a}_A \bar{x} \\ \bar{\sigma}_A^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i^A - \bar{a}_A x_i^A - \bar{b}_A)^2 \end{cases}, \quad (11)$$

with  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i^A$  and  $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i^A$ .

Given the parameters, a test sample  $y_j^A$  follows the following Gaussian distribution

$$y_j^A - \bar{a}_A x_j^A - \bar{b}_A \sim \mathcal{N}\left(0, \bar{\sigma}_A^2 \left(1 + \frac{1}{n} + \frac{(x_j^A - \bar{x})^2}{(n-1)s_x^2}\right)\right),$$

and its confidence interval is of the form

$$y_j^A = \bar{a}_A x_j^A + \bar{b}_A \pm Q(1 - \alpha)\bar{\sigma}_A \left(1 + \frac{1}{n} + \frac{(x_j^A - \bar{x})^2}{(n-1)s_x^2}\right)^{1/2}.$$

Analogously to that presented in Section 2.2, the mean accuracy  $\bar{y}^A$  satisfies  $\bar{y}^A - \bar{a}_A \bar{x}^A - \bar{b}_A \sim \mathcal{N}(0, \hat{\sigma}_A^2)$ , and with probability  $1 - \alpha$  the confidence interval of  $\bar{y}^A$  is

$$\bar{y}^A = \bar{a}_A \bar{x}^A + \bar{b}_A \pm Q(1 - \alpha)\hat{\sigma}_A. \quad (12)$$

To compare the two systems  $A$  and  $B$ , the confidence interval of  $\bar{y}^A - \bar{y}^B$  is now calculated by

$$\bar{y}^A - \bar{y}^B = \bar{a}_A \bar{x}^A - \bar{a}_B \bar{x}^B + \bar{b}_A - \bar{b}_B \pm Q(1 - \alpha)\sqrt{\hat{\sigma}_A^2 + \hat{\sigma}_B^2}. \quad (13)$$

## 3. EXPERIMENTS

Here we summarize the main experiments conducted, which consist of the estimation and comparison of the performances of three pre-trained chord recognition systems: the HP system [8] that is trained on the audio dataset used in the MIREX Chord Detection task 2010<sup>4</sup>, the labROSA

<sup>4</sup>[http://www.music-ir.org/mirex/wiki/2010:Audio\\_Chord\\_Estimation](http://www.music-ir.org/mirex/wiki/2010:Audio_Chord_Estimation)

System	S model [%]	I model [%]	L model [%]	GT acc [%]
HP	$82.6 \pm 1.3$	$82.2 \pm 1.3$	$82.2 \pm 1.3$	82.2
labROSA	$76.6 \pm 1.3$	$77.6 \pm 1.4$	$77.6 \pm 1.3$	77.6
Chordino	$75.2 \pm 1.3$	$76.2 \pm 1.0$	$76.2 \pm 1.0$	76.2
Consensus	$82.3 \pm 1.3$	$82.7 \pm 1.2$	$82.7 \pm 1.2$	82.7

System	S model [%]	I model [%]	L model [%]	GT acc [%]
HP - labROSA	$6.6 \pm 1.9$	$4.6 \pm 1.9$	$4.6 \pm 1.9$	4.6
HP - Chordino	$8.0 \pm 1.7$	$6.0 \pm 1.6$	$6.0 \pm 1.6$	6.0
HP - Consensus	$0.3 \pm 1.8$	$-0.5 \pm 1.8$	$-0.5 \pm 1.8$	-0.5
labROSA - Chordino	$1.3 \pm 0.9$	$1.4 \pm 1.7$	$1.4 \pm 1.6$	1.4

**Table 1.** Upper table: the estimation of performance of HP, labROSA and Chordino on the validation set. The first three columns are the estimated mean GT accuracies using Eq. (4), (8) and (12) respectively, where the confidence level is fixed at 95%. The forth column is the real GT accuracies. Lower table: the comparison of performances between test systems, using Eq. (5), (9), (13) and real GT accuracy differences respectively.

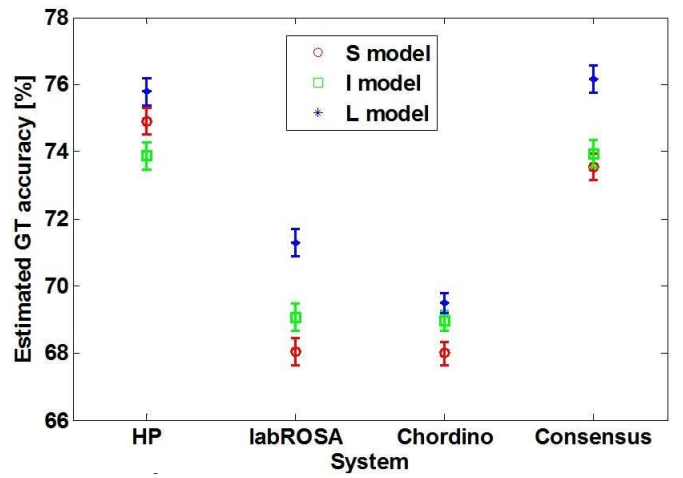
system [2] which is trained on the same dataset, and finally Chordino, a freely-available pre-trained chord recognition system [5]. The reference system used to generate pseudo annotations is the Jump Alignment (JA) method [6], which has shown to produce more accurate chord predictions than all other systems, by means of using the online chord database E-chords. The validation set consists of 175 The Beatles’ songs, of which we have both ground truth and pseudo annotations. This set is used to learn the parameters of the single Gaussian (S), the individual Gaussian (I) and the linear regression (L) models. The test set consists of 1840 songs from a variety of genres, of which we can only derive pseudo annotations using JA. The objective of the experiments is to estimate and compare the GT accuracies of the three systems on the test set, in terms of the S, I and L models.

### 3.1 Verification

We first regarded the validation set as test set so as to verify the confidence intervals estimated by S, I and L models. The results are shown in Table 1. All real GT accuracies fall in the estimated interval with a 95% confidence level, verifying the reliability of the models. We also observed from Table 1 (lower table) that S model biases towards HP as expected, because HP shared the same chromagram features with the JA method. This bias was then removed by using I/L models.

### 3.2 Performance estimation

We then estimated and compared performances of the systems on the large test set, of which the ground truth annotations are not available. Again, we first estimated the mean GT accuracies of the three systems, in terms of S/I/L models and pseudo annotations generated by JA. The results are illustrated in Figure 2. We observed that the estimated accuracies between labROSA and Chordino are highly overlapped, indicating a similar performance of the two systems. Alternatively, there is a large gap between HP and the other two systems, implying the superiority of the HP system. We also observed that S model ranked higher than



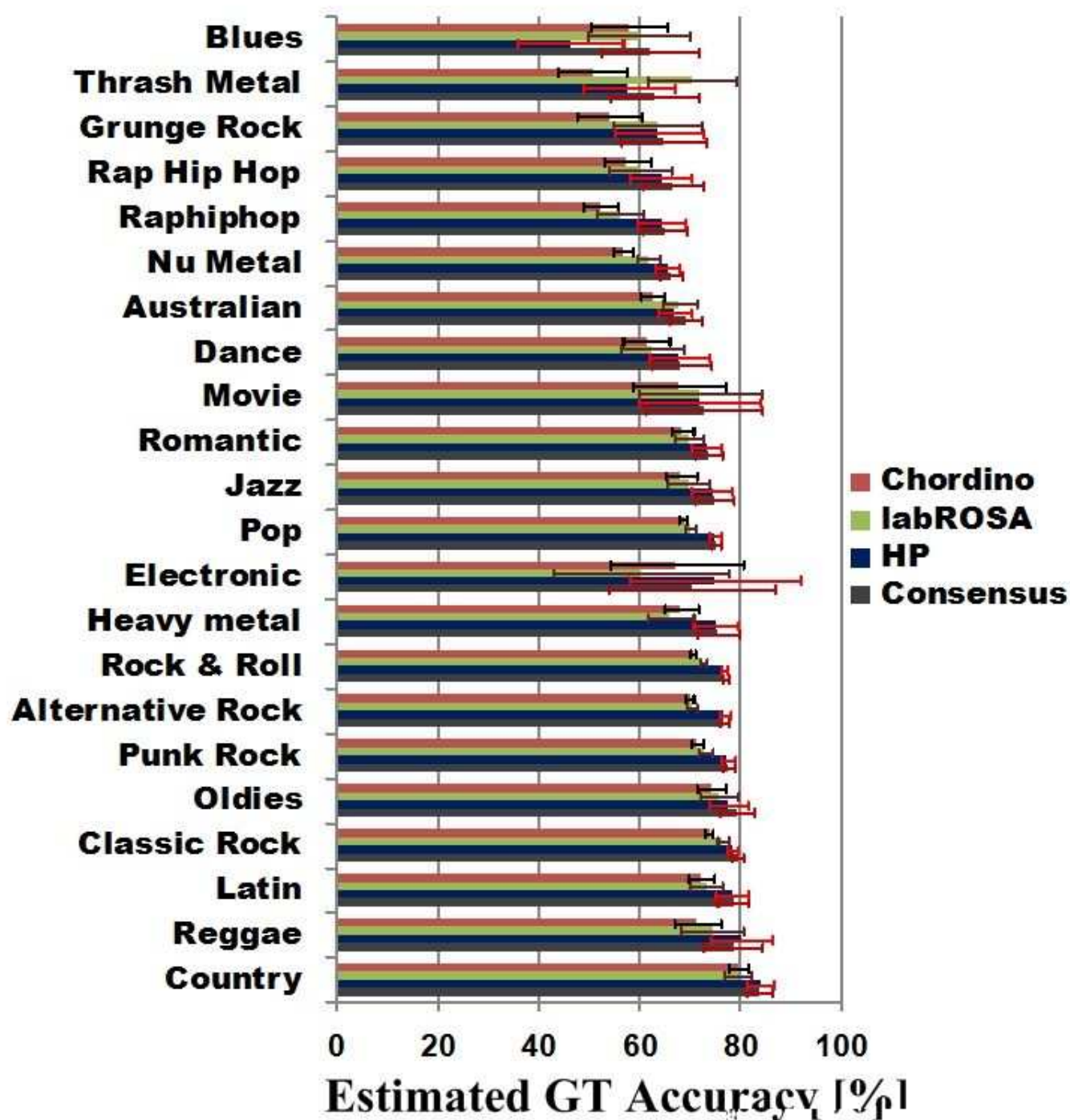
**Figure 2.** The estimated mean GT accuracies of the test systems on a large test set (1840 songs).

I model on HP, which is different from the cases for the other systems. This implies a bias towards HP, which however, was eliminated by I/L models.

Finally, we categorized the test songs by their genre and estimated the mean GT accuracies of the test systems on each genre. The results are illustrated on Figure 3 to Figure 5. We observed that HP performs better on most of genres, especially on Rock related genres. This is consistent with the fact that the system is trained on songs mainly from the Rock genre. Meanwhile, the performances of labROSA and Chordino are highly overlapped, except for some genres containing few songs, which may happen by chance.

### 3.3 Consensus

Inspired by the fact that the best test system HP does not always outperform the other two in genre-specific estimation, we tried to combine predictions from the three systems so as to improve the recognition accuracy. As a trial, we simply combined predictions on each frame by major-



**Figure 3.** Estimated GT accuracies of the test systems on each genre, using S model.

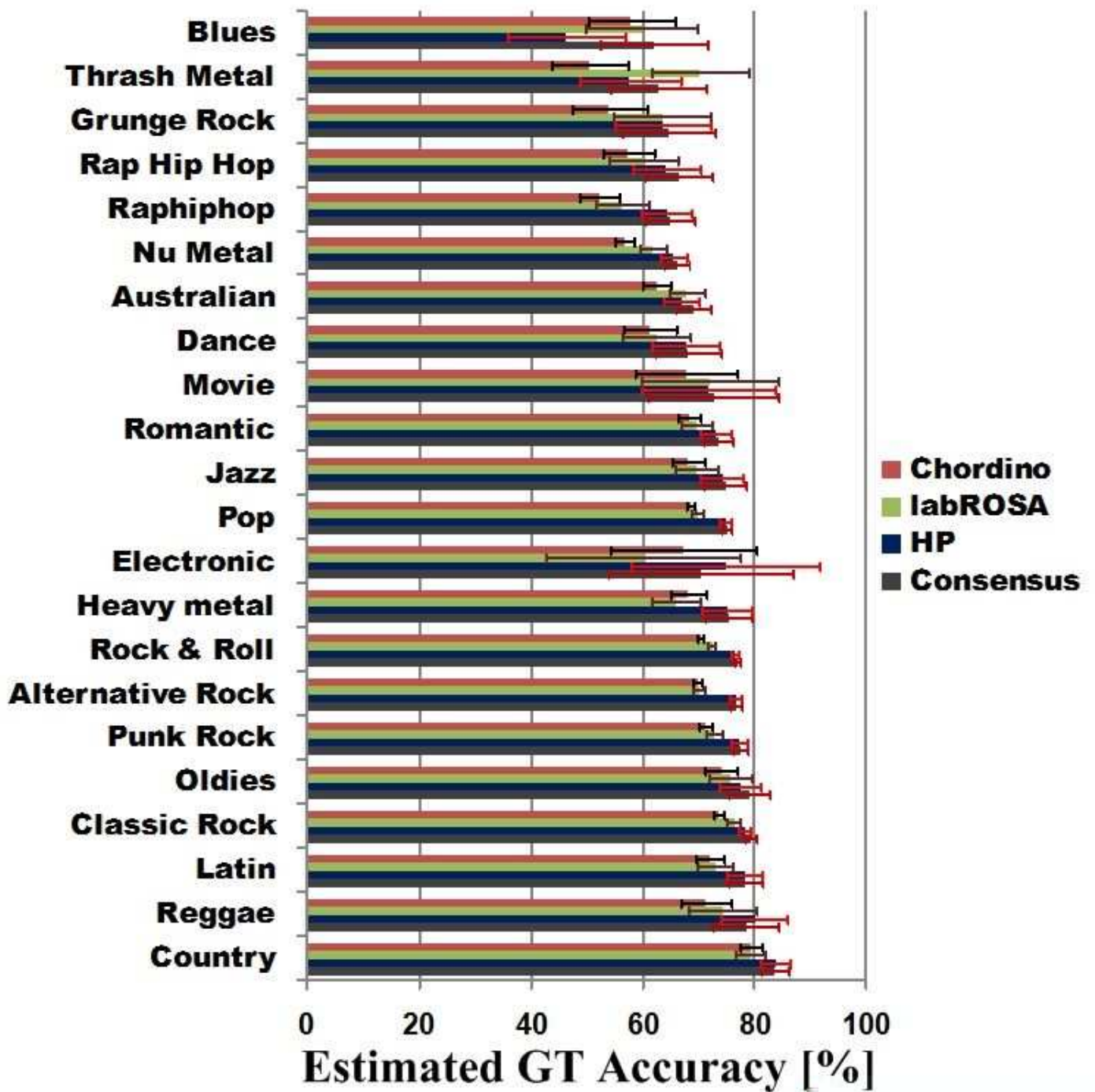


Figure 4. Estimated GT accuracies of the test systems on each genre, using I model.



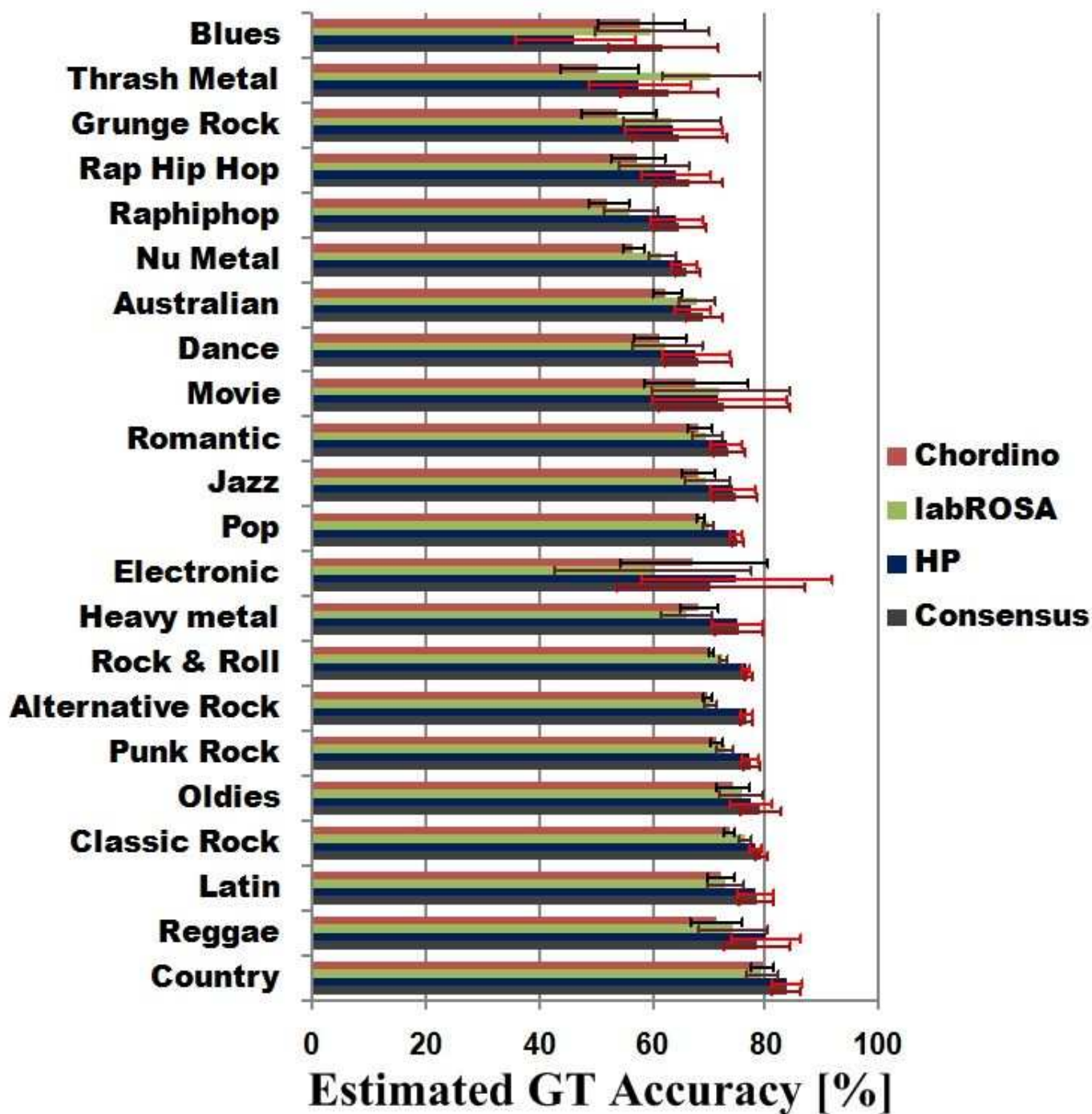


Figure 5. Estimated GT accuracies of the test systems on each genre, using L model.

ity vote, which yielded a consensus prediction of the three systems (denoted by “Consensus”). The performance of the Consensus system are presented in Table 1, as well as Figure 2 to Figure 5. It is promising to see that the consensus prediction performs slightly better than HP (on I/L models), by means of compensating the low performance of HP on certain genres (e.g. Funk and Blues). This observation is sufficiently encouraging that an investigation of combining systems’ predictions using machine learning techniques will be carried out in the future.

#### 4. CONCLUSIONS AND FUTURE WORK

We have proposed a new method to evaluate chord recognition systems on songs which do not have full annotations. The approach goes beyond the existing evaluation metrics, allowing us to carry out extensive analysis on chord recognition systems, such as their generalizations to different genres. In the experiments, we tested this method on three systems, and the resulting confidence intervals on a validation set verified its reliability. We then evaluated these systems on a much larger test set and obtained some promising observations which can not be achieved by current evaluation techniques. These observations inspired us to combine predictions of different systems, and the resulting consensus system achieved the best performance by means of compensating weakness of a specific system.

For the future work, we aim at improving the reliability of the statistical models proposed. Since there may be errors and omissions in chord sequences obtained from the online databases, these chord sequences may become outliers in the validation and test sets (e.g. circled points in Figure 1). A method to detect and remove these outliers is then a direction of our future work. Meanwhile, as pointed out in Section 3.3, an investigation of combining systems’ predictions using machine learning techniques will also be conducted.

#### 5. REFERENCES

- [1] B. Catteau, J. Martens, and M. Leman. A probabilistic framework for audio-based tonal key and chord recognition. In *Proc. of GfKI*, pages 637–644, 2006.
- [2] D. Ellis and A. Weller. The 2010 LABROSA chord recognition system. In *Proc. of ISMIR (MIREX)*, 2010.
- [3] K. Lee and M. Slaney. A unified system for chord transcription and key extraction using hidden markov models. In *Proc. of ISMIR*, 2007.
- [4] M. Mauch. *Automatic chord transcription from audio using computational models of musical context*. PhD thesis, Queen Mary University of London, 2010.
- [5] M. Mauch and S. Dixon. Approximate note transcription for the improved identification of difficult chords. In *Proc. of ISMIR*, 2010.
- [6] M. McVicar, Y. Ni, R. Santos-Rodriguez, and T. De Bie. Using online chord databases to enhance chord recognition. *Journal of New Music Research*, 40(2):139–152, 2011.
- [7] J. Neter, W. Wasserman, and M. H. Kutner. *Applied linear statistical models*. Irwin Press, Boston, 1990.
- [8] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie. An end-to-end machine learning system for harmonic analysis of music. In <http://arxiv.org/abs/1107.4969v1>, 2011.
- [9] H. Papadopoulos and G. Peeters. Local key estimation based on harmonic and metric structures. In *Proc. of DAFX*, 2009.
- [10] T. Rocher, M. Robine, P. Hanna, L. Oudre, Y. Grenier, and C. Févotte. Concurrent estimation of chords and keys from audio. In *Proc. of ISMIR*, 2010.